Chicago Crime Analysis and Prediction

Tejal Bhangale tbkkr@mst.edu
Weerdhawal Chowgule wcmb3@mst.edu
Pooja Zare pzgkq@mst.edu

Abstract

With the proliferation of technology and automation over the years, it is easier to determine the extent of vulnerability an individual is subjected to, at a specific geographic area on any occasion. The main objective of our project is to anticipate if a particular neighborhood in the city, at a given duration of the day will be a crime hotspot or not, with an acceptable rate of accuracy. Our research aims to exploit background criminal knowledge procured from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The CLEAR system is an ideal and reliable source for assimilating criminal occurrences in Chicago, where criminal investigation records are preserved since 2001. The research, with the combination of demographic information, establishes an approach of detecting crimes in a particular geographic area by analyzing and studying the criminal occurrences of the area and thus making deductions by employing well-founded and reliable learning algorithm. The analysis is further extended to incorporate the impact of housing and inhabitation, literacy rate, employment and socioeconomic status on the crime occurrence rate.

1 Introduction

According to the Chicago Crime report records, "Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century. The city's overall crime rate, especially the violent crime rate, is substantially higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US"(Exploring Chicago Crimes 2012-2016 , 2017). Chicago's homicide rate is higher than the larger American cities of New York and Los Angeles, the reasons for the higher numbers in Chicago remain unclear(Crime in Chicago, 2017). However, "the Chicago police department tallies data differently than police in other cities, the FBI often does not accept their crime statistics."

Chicago Crime Report (2012-2016): *In order to record the crime in Chicago, the Chicago police department developed a tool to assist city residents in problem-solving and combating crime and disorder in their neighborhoods, all thing shows Chicago police department has a long history of using data" (Jeanne Clery Disclosure Act, 2015). The report will cover the number of crime, the type of crime and the times series development of crime, etc.* From amazonaws.com: Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century. The city's overall crime rate, especially the violent crime rate, is substantially higher than the US average. Chicago

was responsible for nearly half of 2016's increase in homicides in the US.

Keeping these concerns in mind, we have leveraged the dataset to analyze and perform prediction on crime events (except the murder cases) that occurred in the City of Chicago from 2001 to present. The data resource is from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. For our project, We have more than 6280882 millions of the data records. The research aim to delve deeper into this statistic and deal with missing values in our dataset. We use several graphs to analyze the dataset, and apply predictive analytics to solve some classification and prediction problems on our dataset.

2 Problem Statement

Why is this dataset more important now? Because the number of crime in Chicago increase, more and more people care about their safety, and the government leader want to build a good environment to the citizens. Therefore, it is required to predict the occurrence of a crime at a location at a specific time of a day and to anticipate if a particular neighborhood in the city, at any given duration of the day will be a crime hotspot or not, with an acceptable rate of accuracy. Furthermore, to incorporate the impact of housing and inhabitation, literacy rate, employment and socioeconomic status[1] on the crime occurrence rate.

2.1 Existing solutions

Several crime analysis techniques have been established over the years for identifying and analyzing patterns and trends in crime and disorder. The approach in [5] predicts crime hotspots based on human behavioral data derived from mobile network activity, in combination with demographic information thus considering both people- and place- centric perspectives. The people-centric approach involves individual and collective profiling and further determining behavioral patterns in crimes that are committed by the same offender or group of offenders. While place-centric perspective adopts a distinct modus operandi crime hotspot detection by analysis and the consequent derivation of useful insights leveraging mobile network activity as a source of human behavioral data. [5] establishes an approach called the Series Finder to tackle the problem of detecting precise patterns in crimes that are committed by the same offender or group of offenders. The learning algorithm processes information similarly to how crime analysts process information instinctively: the algorithm searches through the database looking for similarities between crimes in a growing pattern and in the rest of the database, and tries to identify the modus operandi (M.O.) of the particular offender. The M.O. is the set of habits that the offender follows, and is a type of motif used to characterize the pattern. The approach to pattern discovery captures several important aspects of patterns:

- Each M.O. is different.
- General commonalities in M.O. do exist.

- Patterns can be dynamic.

[5] exploited data procured from the Datathon for Social Good - organized by Telef´onica Digital, at Open Data Institute and MIT during the Campus Party Europe 2013, London September 2013. The data provided by participants were categorized into two sections (i)smart-step data: smart anonymized and aggregated human behavioral data computed from mobile network activity in the London Metropolitan Area (ii)geo-localised open data: which included reported criminal cases, residential property sales, transportation, weather, etc. For each Smart-steps cell (geographic lo- cation with precise lat, log values), a prediction was made whether that particular cell will be a crime hotspot or not in the next month.

Recent works in the domain are influenced by the proliferation of social media which has sparked an interest in using data from software applications to anticipate a variety of variables, including electoral outcomes and market trends. Following the trend, Wang et al. [6] proposed the applicability of social media to predict criminal incidents. Their approach relies on a semantic analysis of tweets using natural language processing along with spatio-temporal information derived from neighborhood demographic data and the tweets meta-data. The authors in [3] proposed a model named Series Finder which formulates predictions if an individual or a group of individuals will commit a crime in the near future by studying the past records and analyzing the pattern. Series Finder first uses one of the crimes as "seed" and then links it with the other crimes the criminal is involved using similarities such as type, location, etc., which leads to a pattern. To be more thorough, the crimes that are grouped together on the basis of similar patterns, have some attributes that are almost identical and are calculated by using learned weight of each crime type and pattern based weights. Validation of the model was carried out by using three different patterns (i) an existing(i.e., original) pattern (ii) a predicted pattern (iii) a verified pattern and based on these results success and the failures are recorded. Recent years have witnessed an augmented expansion of social media with millions of users. One such example being the twitter, which is considered to be one of the most abundant resources in the field of varied data. Statistical topic modeling and linguistic analysis of twitter-specific data are used to identify the major discussions, including crime, across various cities.

Major attempts are taken to incorporate these techniques in extracting the crime related discussion across Chicago city in creation of a crime prediction model [4]. Gerber collected all the information between January 1, 2013 and March 31, 2013 documented by the Chicago Crime Department. The research considered the time-stamp of occurrence, latitude/longitude coordinates of the crime at the city-block level, and one among the 27 crime types provided by the Crime department. For the same period they considered twitter data, from official Twitter Streaming API, tagged with GPS coordinates of the city of Chicago [4]. The density of the tweets were correlated with different crime types documented with the crime department to predict the crime occurrence.

Drawbacks: There are many challenges to using Twitter as an information source for crime prediction. Tweets are notorious for (un)intentional misspellings, on-the-fly word invention, symbol use, and syntactic structures that often defy even the simplest computational treatments (e.g., word boundary iden-

tification). These factors conspire to produce a data source that is not only attractive – owing to its real time, personalized content – but also difficult to process. Thus, despite recent advances in all stages of the automatic text processing pipeline (e.g., word boundary identification through semantic analysis) as well as advances in crime prediction techniques (e.g., hot-spot mapping), the answer to the primary research question in [4] has remained unclear.

3 Proposed Approach

Chicago, Illinois ranks third in the United States in population (2.7 million), second in the categories of total murders, robberies, aggravated assaults, property crimes, and burglaries, and first in total motor vehicle thefts [3]. In addition to its large population and high crime rates, Chicago maintains a rich data portal containing, among other things, a complete listing of crimes documented by the Chicago Police Department[2]. Analysts study crime reports, arrests reports, and police calls for service to identify emerging patterns, series, and trends as quickly as possible. With the adoption of automation and machine learning techniques, these activities can be accomplished at a more accelerated and efficient rate. Our objective is to predict the type of criminal activity irrespective of the perpetrator. We have procured criminal records from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Table 1 depicts the attributes considered by our system model and is referred to as the dataset.

To accomplish our objective of recognizing crime patterns across the city based on geographical locations, our first measure is dividing the entire city of Chicago into smaller units called cells, each district of Chicago is evaluated as a cell. In the meta-data obtained from the CLEAR system of Chicago Police Department, each criminal record is characterized by several attributes that includes crime description, location, longitudes and latitudes, etc as elaborated in Table 1. These attributes comprise the dataset of the system model adopted by our project and will be conducive while plotting the exact locations of the crimes. In addition, the CLEAR system classifies the crimes into 32 different categories as depicted in Table 2. With all the attributes, we expect to depict the pattern of each crime-type across the City of Chicago for an entire year. For a sound prediction of the occurrence of a crime at any location and any hour of a day, it is required to consider the data that is consistent and out of exemptions. Therefore in order to abstain from false conjectures and guarantee a reliable prediction model, we plan on considering the criminal records of past 6 years to train our algorithm. The derived prediction model is then tested against the records from recent years for validation and determining the accuracy rate of our model.We build two types of models, binary classification model for predicting whether a crime is severe or not, and multiclass classification model to determine the exact type of crime. To increase the reliability of the prediction we intent to compare the accuracy rate among various machine learning classifiers.

Variable	Description
ID	Unique identifier for the record.
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date	Date when the incident occurred. this is sometimes a best estimate.
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description.
Primary Type	The primary description of the IUCR code.
Description	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Description of the location where the incident occurred
Arrest	Indicates whether an arrest was made.
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
District	Indicates the police district where the incident occurred.
Ward	The ward (City Council district) where the incident occurred.
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas.
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Year	Year the incident occurred.
Updated On	Date and time the record was last updated.
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Table 1: Attributes of crime-data provided by the Chicago Police Department

3.1 Classification

In this project, we perform data modeling using classification models or classifiers. There are mainly two types classification-Binary and Multiclass. Binary classification is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a classification rule. There are many metrics that can be used to measure the performance of a classifier or predictor; different fields have different preferences for specific metrics due to different goals. In machine learning,

Theft	Battery
Robbery	Criminal Damage
Deceptive Practice	Narcotics
Domestic Violence	Non-Criminal (Subject Specified)
Assault	Criminal Trespass
Gambling	Arson
Burglary	Prostitution
Concealed Carry License Violation	Human Trafficking
Motor Vehicle Theft	Weapons Violation
Homicide	Offense involving Children
Crime Sexual Assault	Sex Offense
Obscenity	Non-Criminal
Liquor Law Violation	Interference with Public Officer
Kidnapping	Public Peace Violation
Intimidation	Stalking
Public Indecency	Ritualism

Table 2: Classification of crimes

multiclass or multinomial classification is the problem of classifying instances into one of three or more classes. (Classifying instances into one of the two classes is called binary classification). Here we apply various types of classifiers to model our dataset and compare their accuracies. Following are the methods we applied for classification:

- **Logistic Regression**

  The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor.

- **Linear SVM**

  Support vector machines (SVMs) are a set of supervised learning methods which learn from the dataset and used for classification. Given a set of training examples, each marked as belonging to one of two classes, an SVM algorithm builds a model that predicts whether a new example falls into one class or the other.

- **Decision Tree**

  Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

- **Naive Bayes model**

  Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

- **Random Forest**

  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

- **kNN**

  kNN is considered among the oldest non-parametric classification algorithms. To classify an unknown example, the distance from that example to every other training example is measured. The k smallest distances are identified, and the most represented class by these k nearest neighbors is considered the output class label.

4 Technical details of proposed approach

4.1 Data Preprocessing

In this part, we clean up the crime dataset obtained from the city of Chicago data portal. In our project, we used the most recent data from year of 2011 to 2017 from the Chicago crime dataset, instead of using the whole dataset which contains 6 million records back to year 2001. We chose not to use earlier data because the distribution of crimes can change over time due to various reasons, and the most recent 5 to 6 years will be more relevant to the current situation, since we view the problem as a classification problem instead of a time-series problem.

We cleaned up the data as below:

1. **Excluded records with vague information:** For example "others" or typo or NA. We can see there are a lot of information, but some unnecessary columns are also removed.
2. **Removed extremely rare crime types and locations :** Though we hope to classify every single type of crime, it is not really practical since some crimes are happening at extremely low frequency. Based on the sorted list of crimes, we can see that some crimes are really rare, for example human trafficking. Though these rare records won't greatly affect our model, we only picked top 25 types of crime for simplicity. As we want to build a classification (yes-or-no) model to determine whether a crime is going to be severe or not, we classified the following types of crimes as severe and give them an indicator variable 1 : "Arson", "Assault", "Battery", "Crime Sexual Assault", "Criminal Damage", "Criminal Trespass", "Homicide", "Robbery". In our case, severe means only the crime that involves direct violence, and everyone near the crime scene needs to use caution immediately. Another important feature for classification is "Location Description". We can imagine robbery is more likely to happen on a street compared with a restaurant, while deceptive practice is probably going to happen at a store instead of at home. To avoid the curse of dimensionality , we dropped crimes that happen at extremely unlikely locations, and only kept crimes that happen at the top 25

locations. Even after two rounds of filtering, we still kept around 85% of the crimes.

3. **Converted categorical features to dummy (indicator) variables for classification :** We checked some other features that can be used for the classifier. Police district and community area are the two features that can be important, since we can expect bad neighborhood may have higher chance to have crimes that are more violent. We could see there are 77 community areas, and 22 districts. This is reasonable because one police district can cover multiple areas. In order to do classification, it is preferable to convert the categorical data into dummy variable (0 and 1). We converted crime type, police district, and location description into dummy variables.

4. **Extracted the time information and converted categorical features to dummy variables :** For each crime time record , we segmented the time into eight blocks of 3-hour, for example 0-3am, 3-6am, etc and created indicator variable based on that. In the data cleanup step, we saved both the exact information and the "binned" information to use it later.

5. **Added an extra feature "distance to closest police station" :** We calculated an additional feature using the coordinate info- the distance of the location to the closest police station. As we can imagine, it is not very likely that a severe crime like robbery is going to happen right in front of a police station. Since we have the longitude / latitude data of each crime, we simply calculated distances between the crime scene and all the police stations in Chicago and determined the distance of the crime to the closest police station.

6. **Integrated socioeconomic status of the neighborhood where the crime happened :** Given a location, there are other relevant information such as income, education level and population structure of the neighborhood. We included several columns of continuous variables describing income, housing condition, education level, population structure based on the "Community Area". The data is also provided by the government of Chicago. To avoid the duplication, we used police district and not the community area as a categorical feature.

4.2 Data Exploration

In this part, we visualized the data using different ways. We analyzed trend of crime occurrence for each year (figure 1). Then, we plotted crime occurrence rates of the following: crime type, scene of crime, hour-day-month of crime (figures 2-6). This gives us a better understanding of the major crimes that occur. We interpret that theft has the highest percentage and is the crime type with the highest crime rate. We also identified locations that are more prone to crimes, street being the scene with highest crime rate.

The crime rate with respect to time of day, day of week and month, depicts which crime happens the most and at what time. Observation says that higher crime intensity is between the duration 12:00PM – 18:00PM. To dive deeper into the data, we focused on only the 4 major types of crimes and grouped the rest into others, that gave us 5 broad categories of crime types. We did breakdown the crimes by location of crime scene, hour of the day, weekday and month as depicted (figures 7-10), to see if we can spot any trends. We have plotted normalized crime types for each category.

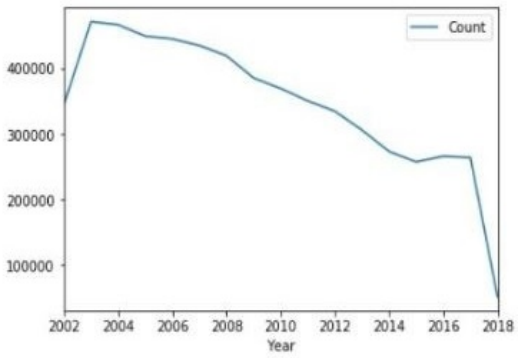

Figure 1: Crime Trend Analysis

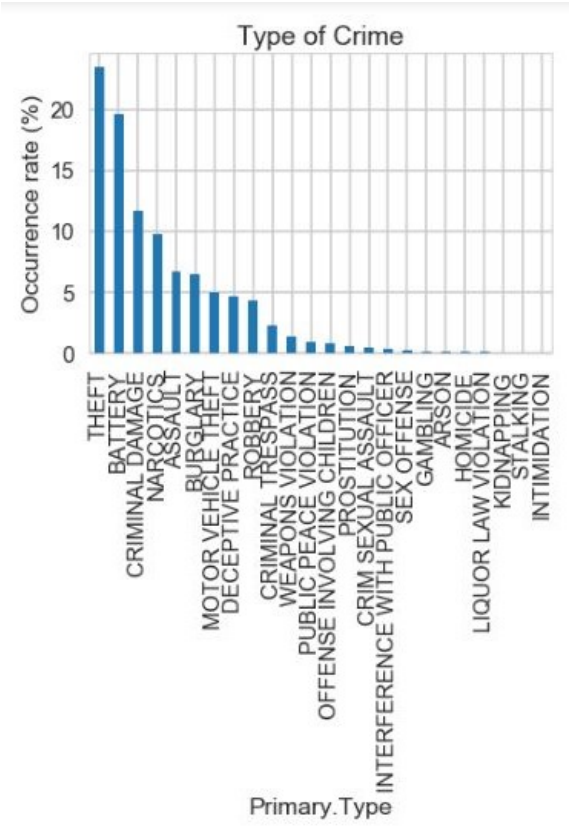

Figure 2: Crime Occurrence Rate V/s Crime Type

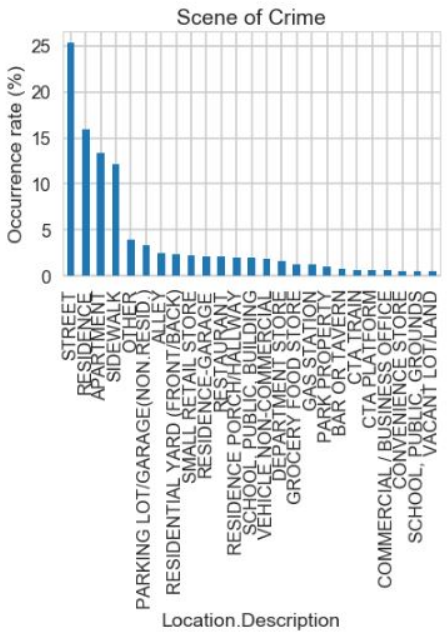

Figure 3: Crime Occurrence Rate V/s Scene of Crime

Figure 4: Crime occurrence rate V/s Time of Crime

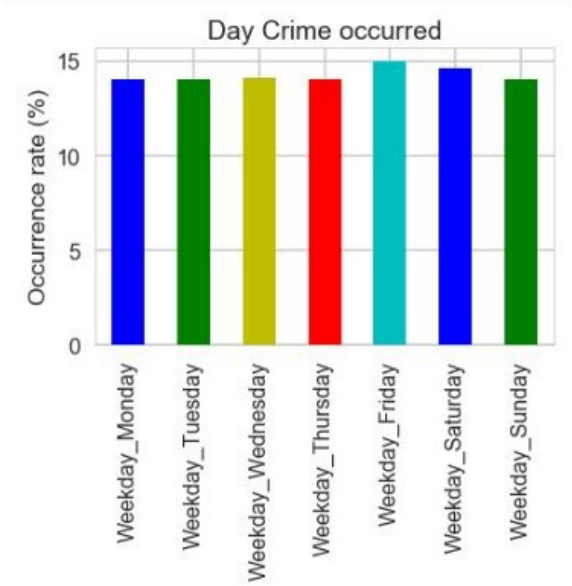

Figure 5: Crime Occurrence Rate V/s Day of Crime

Figure 6: Crime occurrence rate V/s Month of Crime

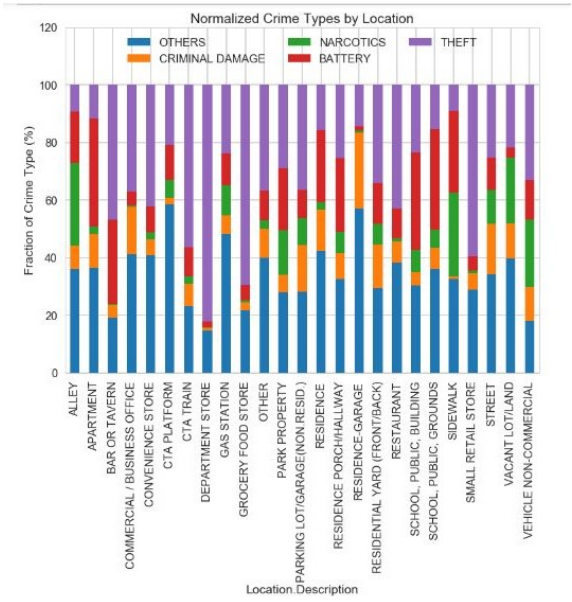

Figure 7: Normalized Crime Types by Location

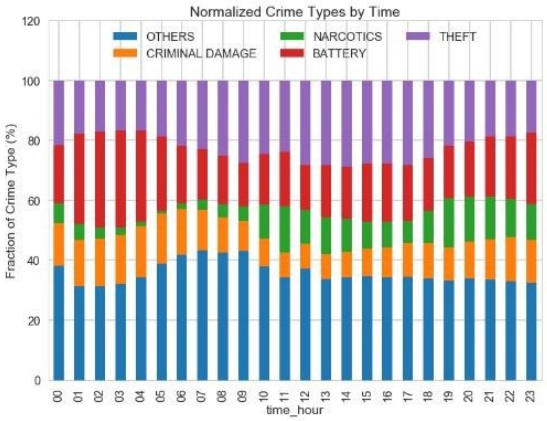

Figure 8: Normalized Crime Types by Time

The density of crime occurrence is demonstrated using a heat map (figure 11). The graph shows the geographical distribution of the crimes in the city of Chicago. The intensity of the color shows the number of crimes. The higher the intensity and higher is the number of crime in that particular area of City of Chicago. This could give a quick insight on which region in the City of Chicago ranks the highest in Crime Rates. The graph is plotted based on the latitude and longitude of the crimes.

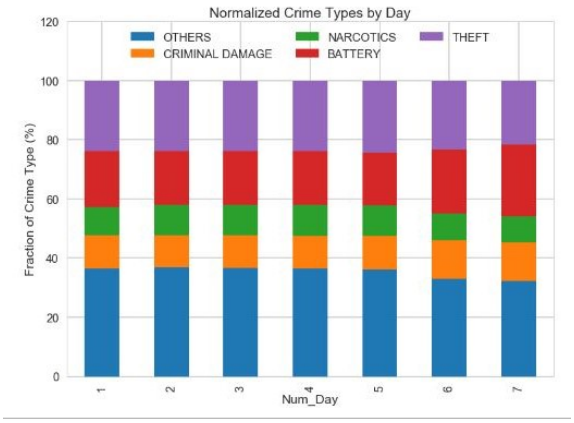

Figure 9: Normalized Crime Types by Day

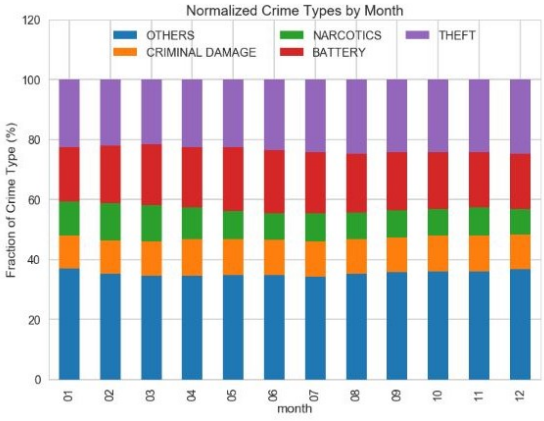

Figure 10: Normalized Crime Types by Month

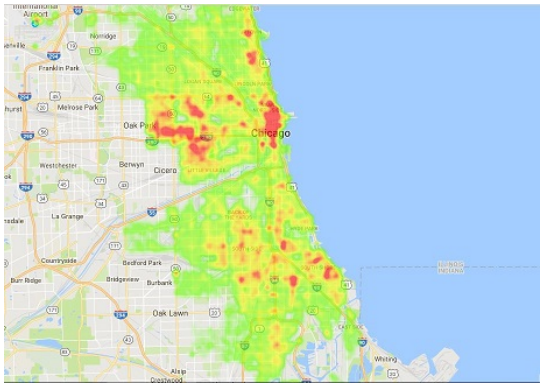

Figure 11: Heat-map of top 5 most dangerous communities w.r.t crime occurrence

4.3 Data Modeling

In this part, we tried to build two types of models, binary classification model for predicting whether a crime is severe or not, and multi-class classification model to determine the exact type of crime.

4.3.1 Binary Classification

We want to predict severity of the crime, so we dropped the exact crime type for that point. These are the features that required normalization: Latitude, Longitude, closest_station. Other features for our dataset are all categorical data and have all been converted to dummy variables (District, Time_block, Weekday, Location description). We performed split on the 80000 sample records, into training set and test set, and normalized by themselves.

Then we performed exploratory data analysis to see what features can be important.

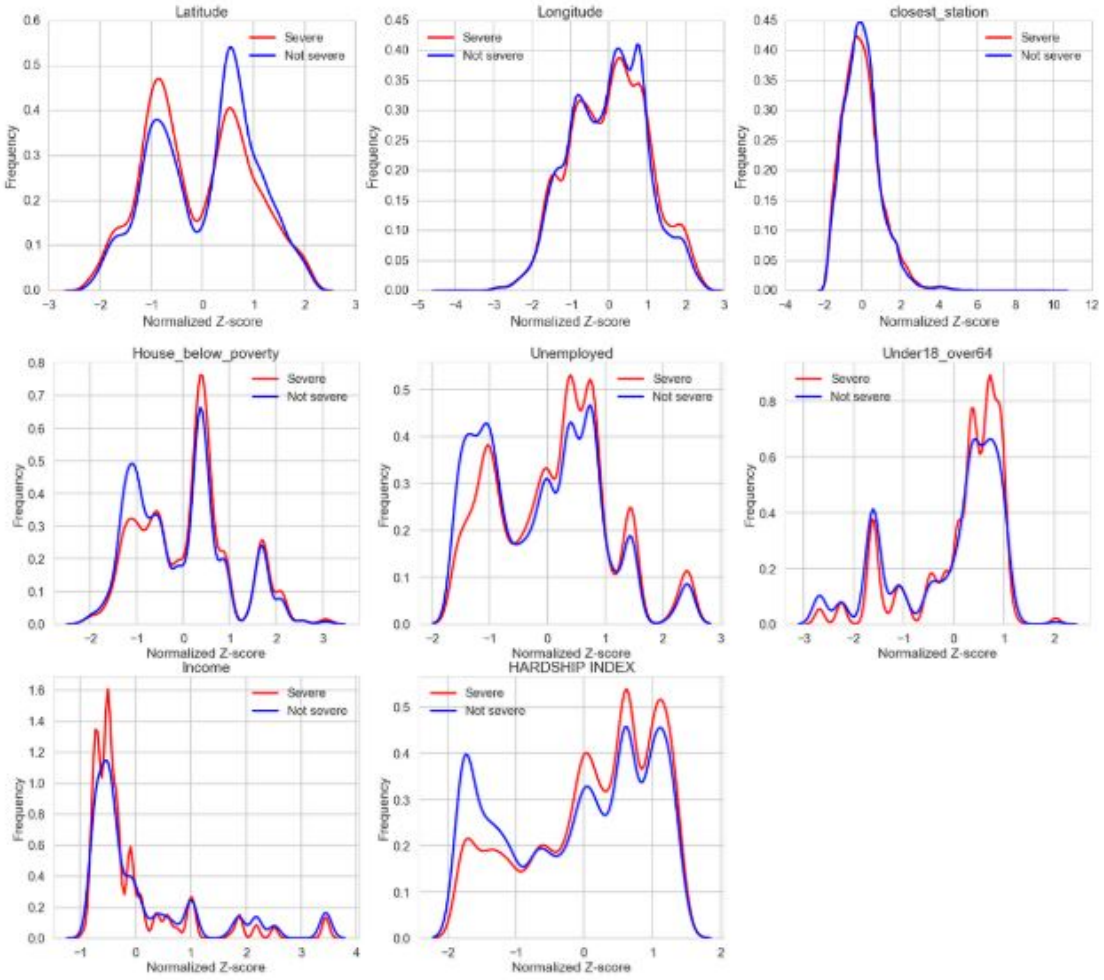

Figure 12: Analysis of Continuous Features

We can see the latitude and longitude info may be helpful for classification. For example, if the latitude is low, it's more likely that severe crime will happen. This is consistent with the fact that the safety at south Chicago is notoriously bad. The economic, unemployment, age status do provide expected result. For example, for regions with lower income and higher unemployment rate, the crimes are going to be more severe.

Then we explored all 60 indicator variables we have to see whether any of them will be good to judge whether a crime is severe or not. From all of those, we could see some indicator variables are good to for classification.

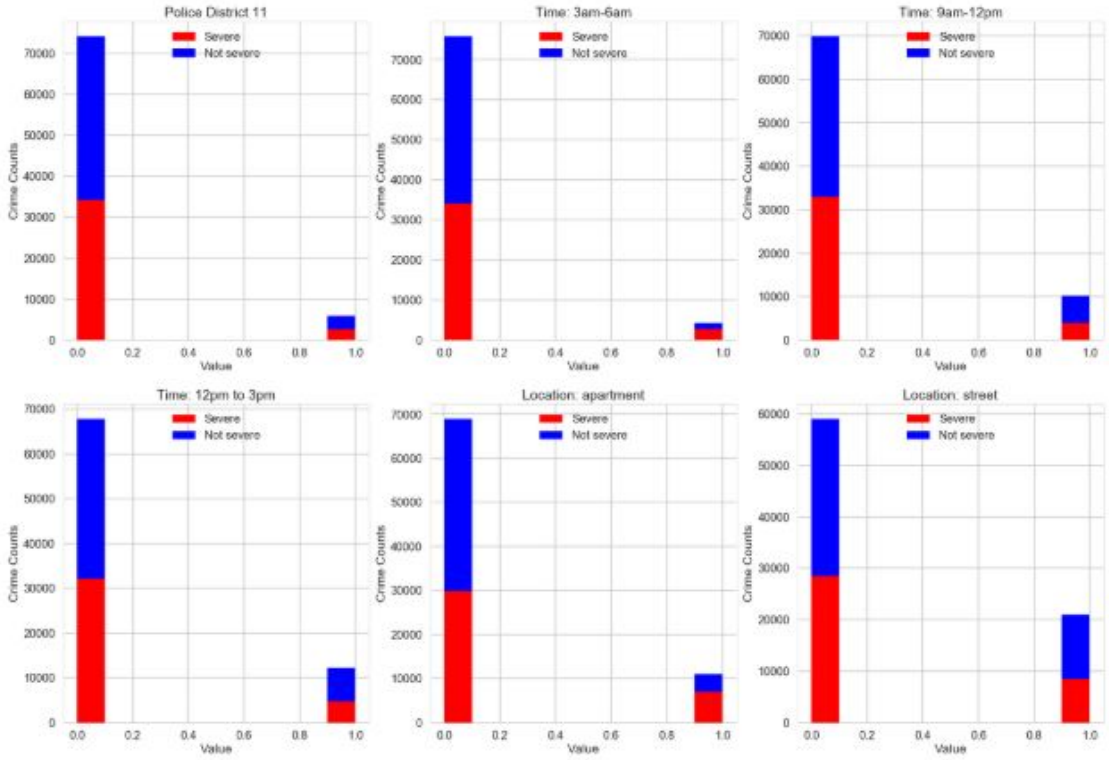

Figure 13: Analysis of Indicator Variables(features)

For example, if the time block is 3am to 6am, the proportion of severe crime is obviously higher, which is not the case in time block 9am to 12pm. If the crime takes place in an apartment or house, it's more likely to be a severe crime, while if the crime takes place on a street where everyone can see, it's less likely going to be severe.

4.3.2 Multiclass Classification

In addition to the above classification, we move down to multi-classification. The goal here is to classify each record into specific type of crimes. We first cleaned up the data as the response variable specifically needs to be transformed into numpy array. Another very important thing is that we have 25 classes of crimes. Our current computational resource does not allow us to do multiclass classification at this scale. Therefore, we here only picked up top four types of crimes. We only maintained these types of crimes: THEFT, BATTERY, CRIMINAL DAMAGE, NARCOTICS, that we are interested in classifying and assigned an integer identifier to them. We split the data into training and testing sets as explained above.

5 Experiments and Final Analysis

5.1 Binary classification

For any classification problem, there is always a Baseline Model: classify everything as the class that occurs most frequently. In our case, the two classes have comparable percentage. The "severe" crime accounts for 46% of the total crime, while the "non_severe" is around 53%. As a result, if we use classification accuracy as the only criteria to judge whether a model is good or not, a good model should at least have the baseline accuracy of 54%. We considered other different classifiers: Logistic regression with Lasso-based feature selection, Linear SVM, Decision tree, Naive Bayes model, Random forest, KNN. We divided our data into a training set and a test set, built different classifiers using the training set, and examined the accuracy of all the classifiers using the test set.

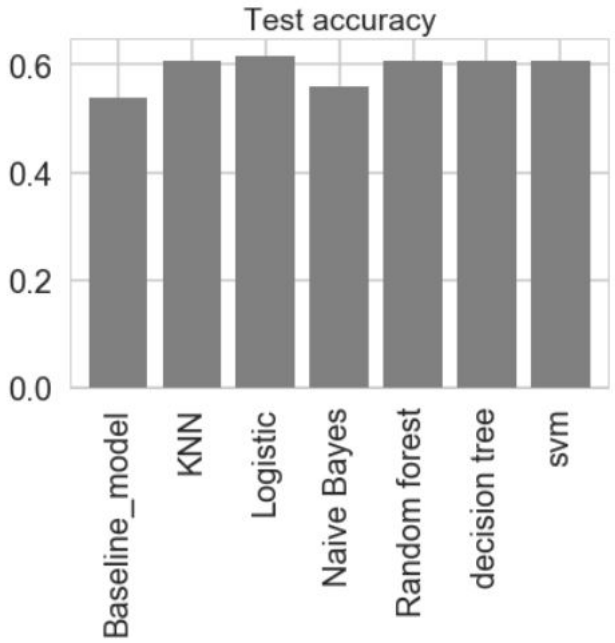

Figure 14: Testing Accuracies of Binary Classification

Based on the accuracy comparison of different models, all the classifiers we tried beat the baseline model, and some perform better than the others. For example, the best classifier is the logistic regression classifier and can reach an accuracy around 63%. The decision tree and random forest also works well. We also identified important factors determining whether a crime is severe or not based on coefficients. The most important features are all indicator variables describing the exact location of the crime, for example if it happened in a department store, it's most likely not going to be something big, and probably will be things like theft. These features are consistent with what we plot above before we set up the model.

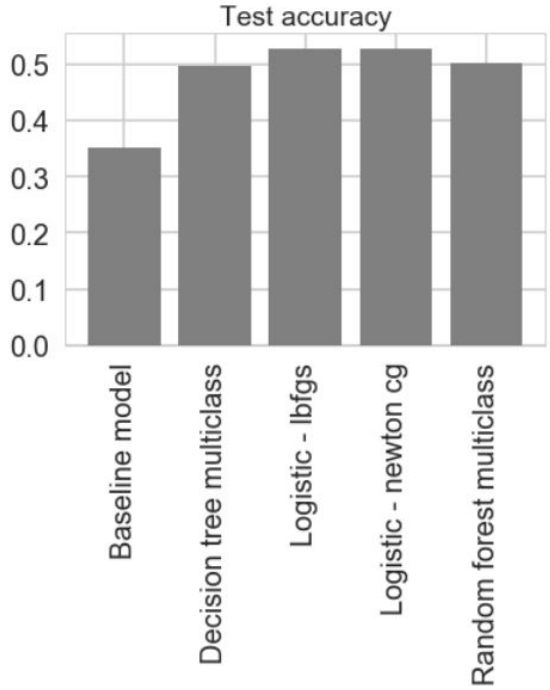

Figure 15: Testing Accuracies of Multi-class Classification

5.2 Multiclass classification

We also built above classification models along with Baseline model for multiclass classification using the train set and compared their accuracies test dataset. As we know that theft is the most common type of crime, what if we predict every single crime record as theft? This acted as our baseline model, which gave accuracy of 33%. To reduce the computational effort, we tried to classify crime records into four major types of crimes: theft, battery, criminal damage and narcotics. From the results above, we can see that all our models outbeat the baseline model for 15% to 20% and logistic regression classifier reaches a final accuracy for more than 50%. We also identified features important for multiclass classification.

6 Conclusion

In our exploratory data analysis, we revealed that several features, for example location of the crime scene and the time of the crime, are associated with the type of crime, providing the basis for out modeling later. For example, crimes happening in department store are more likely to be thefts. Crimes happening at late hours are prone to be violent in nature. One single feature may not be sufficient for determining the type of crime, but a combination of various features can be powerful. This is indeed the case in our modeling. We have also taken into account Socioeconomic status of Chicago city.

In our modeling, we started with binary classification to determine whether a crime is going to be a severe crime or not, and the result suggests that our best models, logistic regression, decision tree and random forest, can reach an accuracy around 63%, beating the baseline model (predicting everything as

non-severe) for 8%. The advantage of our model over the baseline model is not huge, because severe or not is a definition that is not 100% clear. However, our first try has demonstrated that integrating different information, we can do classification of crimes. Additionally, we also revealed features that contribute most to determining whether a crime is severe or not.

In addition to binary classification, we also tried multiclass classification to determine whether we can accurately classify a crime record into one of the four major types of crime: battery, theft, criminal damage and narcotics. The result showed that our best multiclass model, logistic regression, reaches a classification accuracy around 50% and beats the baseline model (around 34% accuracy).

As for future work, in order to boost the classification accuracy, it will be necessary to incorporate other information. For example, the police department may focus on solving a specific type of crime during a specific period of time, which may reduce the occurrence of that type of crime. Additionally, some events and the outcomes of the events may be associated with some crime types, for example basketball games, baseball games and elections. Weather information and classification of buildings can also be incorporated. It will be interesting to see whether these other features can help the classification.

References

[1] Chicago data portal - hardship index https://data.cityofchicago.org/health-human-services/hardship-index/792q-4jtu.

[2] Clear dataset by chicago police department https://data.cityofchicago.org/public-safety/crimes-2001-to-present/ijzp-q8t2.

[3] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: a general framework and some examples. *computer*, 37(4):50–56, 2004.

[4] Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[5] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer, 2013.

[6] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. *SBP*, 12:231–238, 2012.